Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 151 (2019) 1207–1212

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Web Search and Data Mining (WSDM) April 29 - May 2, 2019, Leuven, Belgium

Association Rules Extraction for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm

Jesus Silva[a]*, Noel Varela[b], Luz Adriana Borrero López[c], Rafael Humberto Rojas Millán[d]

[a] *Universidad Peruana de Ciencias Aplicadas, Lima 07001, Peru*
[b,c,d] *Universidad de la Costa (CUC), Barranquilla 080003, Colombia*

Abstract

Data Mining applied to the field of commercialization allows, among other aspects, to discover patterns of behavior in clients, which companies can use to create marketing strategies addressed to their different types of clients. This research focused on a database, the CRISP-DM methodology was applied for the Data Mining process. The database used was that corresponding to the sector of SMEs and referring to customers and sales, the analysis was made based on the PFM model (Presence, Frequency, Monetary Value), and on this model the grouping algorithms were applied: k -means, k-medoids, and SelfOrganizing Maps (SOM). To validate the result of the grouping algorithms and select the one that provides the best quality groups, the cascade evaluation technique has been used applying a classification algorithm. Finally, the Apriori algorithm was used to find associations between products for each group of customers.

Keywords: marketing; SMEs; data mining; CRISP-DM methodology; Association Rules Extraction; Apriori algorithm.

1. Introduction

Companies generate and store a large amount of information every day, but the data, as it is stored, does not usually provide direct benefits, its real value lies in the information we can extract from them, that is, information

* Corresponding author. Tel.: +51920287620
E-mail address: jesussilvaUCP@gmail.com

10.1016/j.procs.2019.04.173

that helps to make decisions or improve the understanding of the phenomena that surround it (Al-hagery, M., 2015) [1]. In this context, Data Mining (DM) allows extracting sensitive information implicitly contained within the data. The implementation of DM processes has allowed to determine the accounting characteristics of the most profitable companies, as well as the profile of their clients, through segmentation (Amelec, V. 2015) [2], (Fernandez-bassso, C; 2016) [3]. Customer segmentation is used as a marketing differentiation tool, which allows organizations to understand their clients and build differentiated strategies (Lis-Gutiérrez M. et al; 2018) [4], (Kulkarni, A. and Mundhe, D., 2017) [5]. Based on what has been described, the purpose of this research is to obtain the client´s segmentation on a sample of companies belonging to the SMEs sector in Colombia through the application of Data Mining techniques.

2. Literature Review

2.1 RFM analysis

The RFM analysis (Recency, Frequency, Monetary) is a marketing technique used for the analysis of customer behavior (Anuradha, K., Kumar, K., 2013) [6]. This is achieved by examining what the customer has purchased, using three factors: (R) Purchase recency, (F) Frequency of purchase, and (M) Amount of purchase in monetary terms. According to theories and researchers, customers who spend more money or buy more frequently in their company, are those clients who end up being more sensitive to the information and messages that the company is transmitting. Similarly, customers who recently purchased react better to marketing initiatives than those who have not recently bought. The RFM analysis is based on the well-known "Pareto Law" or the 80/20 , Pareto observed that in his country 80% of the land was owned by 20% of the population. This principle is having a wide spectrum, being applied to plenty of situations. In the case of the RFM analysis, "80% of purchases come from 20% of customers" , that is, 20% of customers generate 80% of sales". This results, even though they seem to be unrealistic, are verified in most businesses and other situations, including sports. This is due to the widespread application of the this law (Hwang, Y, et al; 2015) [7]. When the RFM analysis is used, each client is assigned a rank or category from 1 to 5, in order to qualify them according to the indicated factors. The three indicators together are called RFM "cells". The customer database is analyzed to determine which customers have been "the best customers" in a given period. Customers who have a "5-5-5" range are ideal customers (Vidhate, D., 2014) [8].

2.2 Data mining methodologies

2.2.1 SEMMA

SAS Institute, developer of this methodology, defines it as the process of selection, exploration and modeling of large amounts of data to discover unknown business patterns. The name of this terminology is the acronym corresponding to five basic phases of the process: Sample (Sampling), Explore (Exploration), Modify (Modification), Model (Modeling), Asess (Evaluation) (Larose, D and Larose, C., 2014) [9].

2.2.2 CRISP-DM

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is a method that has proven to guide the Data Mining works. It was created by the group of companies SPSS, NCR and Daimer Chrysler in the year 2000, and is currently the most used reference guide in the development of Data Mining projects (Pickrahn, I. et al; 2017) [10]. It structures the process in 6 phases: Understanding the business, Understanding the data, Preparing the data, Modeling, Evaluation, and Implementation. The succession of phases is not necessarily rigid. Each phase is broken down into several general tasks of second level.

2.2.3 Comparison between the SEMMA and CRISP-DM Methodologies

Table 1 presents a comparison between the SEMMA and CRISP-DM methodologies.

Table 1. Comparison between SEMMA and CRISP-DM Methodologies

Method	SEMMA	CRISP-DM
Phases	1. Extraction of the sample population. 2. Exploration of information. 3. Data manipulation. 4. Analysis and modeling of data. 5. Valuation of results.	1. Understanding of the business 2. Understanding of the data. 3. Preparation of the data. 4. Modeling. 5. Evaluation. 6. Implementation.
License	Linked to SAS products	Free

3. Data and Methods

The data was provided by the Chamber of Commerce of Barranquilla, Colombia, and corresponds to customer records and sales taken from 2015 to 2018 for a group of companies belonging to the SMEs sector in Colombia, Caribbean region [11].

The data collected were categorized as follows:

- Clients: covers the personal data of clients, provides geographical and demographic descriptors, such as ID, RUC, address, age, gender, marital status, telephone, e-mail, workplace, profession, etc.
- Sales: this category has the daily billing records for sale, which provide us with the description of each purchase made by customers during the period 2015-2018.

The databases that contain the data of interest for the analysis are the following (Prajapati, D et al, 2017) [12]:

- Clients: contains personal information of the company's clients. It has a total of 44,800 customer records.
- Client Type: Contains three records representing end customers, distributors and franchisees.
- Institution: defines if the client belongs to a public institution/company, to a private company or is a natural person.
- Invoice: all the billing information recorded by the company during the study period. It has a total of 136278 invoice records.
- Invoice Details: the products that have been purchased in each invoice. It has a total of 403159 invoice detail records.
- Products: contains the records of all the products that the company sells. It has a total of 11127 product records.
- Product Groups: the groups or categories to which the products belong. It has a total of 58 product categories.
- Brands: the brands of the products marketed by the company. It has a total of 396 trademark registrations.

The method used to carry out the process of Data Mining was CRISPDM (Cross-Industry Standard Process for Data Mining) (Abdullah, M and Al-Hagery, H., 2016) [13], (Varela Izquierdo N. et al; 2018) [14], it consists of five phases: Sample (Sampling), Explore (Exploration), Modify (Modification), Model (Modeling), Asses (Valuation), each of them covers a set of activities, which must be followed to carry out a mining process with high quality results.

For the segmentation of Master PC clients based on purchasing behavior, the normalized variables will be taken into account: Receipt, Frequency, and Amount. Taking into account that there is a wide range of clustering algorithms, an analysis has been performed on some of them, corresponding to the most used in this type of case. This allowed selecting the segmentation algorithms that will be applied in the present research. These are: self-

organized maps of Kohonen (SOM), K-means, and CLARA algorithm (cluster for large applications) which is an extension of the k-medoids algorithm (Ban, T et al; 2015) [15].

For the selection of the number of groups, the techniques of internal evaluation will apply sum of error squared and the silhouette index (Witten, I and Frank, E., 2002) [16]. After applying the segmentation algorithms, it will be determined which of them provide the best results based on the described cascade evaluation method in (Vo, B and Le, B., 2009) [17].

4. Results

Each segmentation algorithm applied to the data resulted in customer groups based on the RFM attributes. The cluster number was transformed to a label that describes the level of loyalty of the clients, this was taken as the decision attribute for the generation of classification rules, which were carried out in order to evaluate the cluster methods used (Kamatkar S. et al; 2018) [18], (Khanali, H., 2017)[19]. Table 2 presents the precision results obtained by the LEM2 algorithm.

Table 2. comparison of results for the algorithms KMEANS, K-MEDOIDS and SOM

Methods	Precision
K-means (5 groups)	0.99991
K-medoids (4 groups)	0.99999
SOM (5 groups)	0.99992

Based on the results about the evaluation parameters presented in Table 2, it was determined that the most appropriate method for customer segmentation in the study sample, on the RFM attributes, is the CLARA algorithm, which belongs to the group of k-medoids methods, the following loyalty levels were discovered: Group 1 High, Group 2 Low, Group 3 Medium, and Group 4 Very Low. To generate the association rules, the Apriori algorithm the most commonly used algorithm for the generation of these rules was applied (Shorman, H and Al Jbara, Y; 2017) [20].

The Apriori algorithm is a method to discover sets of frequent elements and generates association rules on a set of transaction data. It first identifies the frequent individual elements through the transactions and then extends to the increasingly large element sets until the resulting element sets reach a specified frequency threshold (support). This algorithm is implemented within the Arules package (Shorman, H and Al Jbara, Y; 2017) [20] of R, some of the generated rules can be seen in Table 3, which describes the specific RFM characteristics for each group of clients.

Table 3. Rules for Customer Profile

Rule	RFM characteristics of clients
1	If a customer has a Frequency of purchase greater than or equal to 5 and a Amount of purchases greater than 16 and less than or equal to 33 dollars, or greater than 92 dollars, then it belongs to the High Loyalty group.
7	If a client has a Recency between 0 and 823 days, and an Amount in purchases less than or equal to 16 dollars, then it belongs to the Low Loyalty group.
10	If a client has a Receipt greater than or equal to 824 days, and a Frequency of purchase less than or equal to 2, and an amount in purchases greater than 500 dollars, then it belongs to the Medium Loyalty group.
16	If a client has a Receipt greater than or equal to 824 days, and a Frequency of purchase less than or equal to 2, and an Amount in purchases less than or equal to 92 dollars, then it belongs to the Very Low Loyalty group.

The results of the application of these rules can be seen in Table 4, regarding the profile of loyalty groups.

Table 4. Loyalty Groups Profile

Group	RFM punctuation			Characteristic
	R	F	M	
High	4	3	4	Clients belonging to this group have a high level of loyalty, have a high level of recency, meaning that their last purchase was made a short time ago, an average of 1 year ago, it also presents a level of frequency between medium and high, that is, they have bought several times, an average of once a year, and a high amount that indicates that they have invested a lot of money in their purchases, an average of 982 dollars.
Medium	3	1	5	Customers belonging to this group have a Medium level of loyalty, made their last purchase some time ago, an average of two years and two months, the number of purchases made on average is one time, but have a very high average purchase amount of 830 dollars. This indicates that they have invested a lot of money in their purchases, considering that they have a low purchase frequency.
Low	4	1	2	Clients belonging to this group have a Low loyalty level, have a high Recency level, that is, they have made their last purchase a short time ago, an average of 1 year ago, but the average number of times they have purchased is 1, and the average amount spent is 22.3 dollars, which indicates that they have invested little money in their purchases. Customers of this group could also be considered as new customers.
Very low	1	1	2	Customers belonging to this group have a Very Low loyalty level, have a very low Recency, which indicates that they have made their last purchase a long time ago, an average of 3.6 years ago, they also have an average frequency of a single purchase, and a Low amount that indicates that they have invested little money in their purchases, an average of 38.6 dollars. Customers of this group could be considered as almost lost customers.

It is important to mention that between the Low and Very Low loyalty levels, more than 50% of the clients are distributed. The marketing experts should use these groups in the way they deem appropriate, for example, could use the strategies of rewarding their best customers to maintain their loyalty, create promotions to attract customers who are at a low level of loyalty, or also offer special discounts to encourage their regular buyers to increase their monetary value, etc.

5. Conclusions

To assess the accuracy of the used algorithms, k-means, k-medoids, and Self Organizing Maps (SOM), classification rules were generated taking, as a decision attribute, the groups created by the algorithms mentioned in this research. Besides, based on the Prediction level, the results suggest that the classification of the groups generated by the CLARA of k-medoids algorithm provide a higher accuracy. The groups of clients of companies belonging to the SMEs sector obtained through the application of Data Mining techniques revealed the loyalty levels: High, Medium, Low and Very Low. These results will allow the company to develop retention strategies towards its customers. The application of the Apriori association algorithm on the set of transactions of each group of clients allowed the elaboration of important association rules with high confidence levels.

References

[1] Al-hagery, M.A., 2015. Knowledge Discovery in the Data Sets of Hepatitis Disease for Diagnosis and Prediction to Support and Serve Community. Int. J. Comput. Electron. Res. 4, 118–125.

[2] Amelec, V. (2015). Increased efficiency in a company of development of technological solutions in the areas commercial and of consultancy. Advanced Science Letters, 21(5), 1406-1408.

[3] Fernandez-bassso, C., Ruiz, M.D., Martin-bautista, M.J., 2016. Extraction of Fuzzy association rules using Big Data technologies 11, 178–185. https://doi.org/10.2495/DNE-V11-N3-178-185

[4] Lis-Gutiérrez M., Gaitán-Angulo M., Balaguera MI., Viloria A., Santander-Abril JE. (2018) Use of the Industrial Property System for New Creations in Colombia: A Departmental Analysis (2000–2016). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[5] Kulkarni, A.R., Mundhe, D.S.D., 2017. Data Mining Technique: An Implementation of Association Rule Mining in Healthcare. Iarjset 4, 62–65. https://doi.org/10.17148/IARJSET.2017.4710

[6] Anuradha, K., Kumar, K.A., 2013. An E-Commerce application for Presuming Missing Items 4, 2636–2640.

[7] Hwang, Y.M., Moon, J., Yoo, S., 2015. Developing A RFID-based food traceability system in Korea Ginseng Industry: Focused on the business process reengineering. Int. J. Control Autom. 8, 397–406. https://doi.org/10.14257/ijca.2015.8.4.36

[8] Vidhate, D., 2014. To improve Association Rule Mining using New Technique : Multilevel Relationship Algorithm towards Cooperative Learning 241–246.

[9] Larose, D.T., Larose, C.D., 2014. Discovering Knowledge in Data. https://doi.org/10.1002/9781118874059

[10] Pickrahn, I., Kreindl, G., Müller, E., Dunkelmann, B., Zahrer, W., Cemper-Kiesslich, J., Neuhuber, F., 2017. Contamination incidents in the pre-analytical phase of forensic DNA analysis in Austria—Statistics of 17 years. Forensic Sci. Int. Genet. 31, 12–18. https://doi.org/10.1016/j.fsigen.2017.07.012

[11] DANE. 2018. Documento metodológico encuesta de desarrollo e innovación tecnológica en la industria Manufacturera. Bogotá: DANE. 43p.

[12] Prajapati, D.J., Garg, S., Chauhan, N.C., 2017. Interesting Association Rule Mining with Consistent and Inconsistent Rule Detection from Big Sales Data in Distributed Environment. Futur. Comput. Informatics J. 2, 19–30. https://doi.org/10.1016/j.fcij.2017.04.003

[13] Abdullah, M., Al-Hagery, H., 2016. Classifiers' Accuracy Based on Breast Cancer Medical Data and Data Mining Techniques. Int. J. Adv. Biotechnol. Res. 7, 976–2612.

[14] Varela Izquierdo N., Cabrera H.R., Lopez Carvajal G., Viloria A., Gaitán Angulo M., Henry MA. (2018) Methodology for the Reduction and Integration of Data in the Performance Measurement of Industries Cement Plants. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[15] Ban, T., Eto, M., Guo, S., Inoue, D., Nakao, K., Huang, R., 2015. A study on association rule mining of darknet big data. Int. Jt. Conf. Neural Networks 1–7. https://doi.org/10.1109/IJCNN.2015.7280818.

[16] Witten, I.H., Frank, E., 2002. Data mining. ACM SIGMOD Rec. 31, 76. https://doi.org/10.1145/507338.507355

[17] Vo, B., Le, B., 2009. Fast Algorithm for Mining Generalized Association Rules 2, 1–12.

[18] Kamatkar S.J., Tayade A., Viloria A., Hernández-Chacín A. (2018) Application of Classification Technique of Data Mining for Employee Management System. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[19] Khanali, H., 2017. A Survey on Improved Algorithms for Mining Association Rules 165, 8887.

[20] Shorman, H.M. Al, Jbara, Y.H., 2017. An Improved Association Rule Mining Algorithm Based on Apriori and Ant Colony approaches 07, 18–23.